POSITION PAPER SUBMISSION TO FOAF2004 WORKSHOP

Estimating whether partial FOAF descriptions describe the same individual

Simon Price, Simon Rawles and Peter Flach
Department of Computer Science, University of Bristol, Bristol BS8 1HH, UK
{simon.price, simon.rawles, peter.flach}@bristol.ac.uk
http://www.cs.bris.ac.uk

Abstract

Accurately determining whether two Friend Of A Friend (FOAF) descriptions describe the same person is crucial to the success of FOAF on the Semantic Web. Unless this is possible, FOAF descriptions of the same individual originating from disparate sources can not be merged together, or "smushed", into a single description. Without smushing, data remains locked within the silos of the originating FOAF descriptions, the holistic view is lost and less can logically be stated about any specific individual. Current approaches to smushing correctly recognize that people do not have a URI unique identifier and instead rely on Inverse Functional Properties (IFP), such as an email address, to act as a weaker proxy for a globally unique identifier upon which to join descriptions. In this paper we point out the inadequacy of this approach in a broad class of FOAF application areas relating to data mining and information extraction. We describe a method capable of smushing numerous automatically extracted partial FOAF descriptions from disparate sources where suitable IFPs are unavailable.

Introduction

The initial wave of Friend Of A Friend (FOAF) [1] documents published on the Semantic Web [2] were largely authored by the person described in the document's Resource Description Framework (RDF) metadata [3]. To a large extent, each person's own FOAF document (or documents in some cases) served as the definitive self-description of the document's author. Importantly, these purpose-written FOAF documents almost universally contain the necessary Inverse Functional Properties (IFP) [4], such as an email address, upon which different RDF descriptions of an individual may be merged. So for this initial wave of FOAF data, leaving aside issues of trust and authenticity, smushing [5] is as trivial as a relational database join on key fields.

Since the initial wave of FOAF self publishing it has become increasingly common for FOAF data to be created automatically from traditional databases (e.g. PARIP membership database [6]) or as part of a hosted web application (e.g. TypePad Weblog service [7]). FOAF from sources such as these also tends to have IFPs suitable for smushing. However, this is not necessarily the case for what we expect to be the next wave of automatic FOAF generation using data mining and information extraction software agents and services. A broad class of FOAF applications is possible using machine learning and data mining techniques to extract information about people from web pages, newsgroups, instant messaging, emails, documents and other electronic resources. Separate FOAF descriptions of the same individual, originating from such diverse sources, are likely to be incomplete and only partially overlapping - frequently without a convenient IFP in common. Consequently, smushing these generated FOAF descriptions is clearly non-trivial and must rely on a range of properties that are neither inverse functional nor unique identifiers. Any solution to this problem must deal with uncertainty and be capable of expressing degrees of confidence in whether two descriptions refer to the same individual.

Overview of Our Approach

Our approach defines a similarity metric across the space of FOAF descriptions and uses this in conjunction with a threshold to determine whether two descriptions should be merged. The overall similarity metric is itself a weighted linear combination of simpler similarity metrics that compare specific combinations of:

- literal values (e.g. *name* and *phone*)
- URIs referenced (e.g. *mbox* IFP if available and *interest*)
- derived values (e.g. initials from *name*, top-level domain from *mbox*)
- relationships between objects (e.g. the *knows* and *seeAlso* graphs)
- properties and content of external documents retrieved over the Web (e.g. a web page made by the person)
- metadata about the FOAF document (e.g. URL, date and referrer)

A good proportion of these metrics go way beyond what can logically be concluded from the strict interpretation of the RDF and OWL semantics of FOAF. For instance, in calculating derived values we treat literal values as non-atomic and also attempt to interpret most URIs as URLs of online resources rather than as simple identifiers. Our metrics are calculated by hand-crafted, domain-specific heuristics modelled on our own intuition of how we might look for clues about identity in the absence of IFPs. For example, given two FOAF descriptions d1 and d2, it is not legal to conclude that a foaf:mbox literal of "mailto:simon.price@bristol.ac.uk" in d1 is quite probably co-referent with the person in d2 with a foaf:name of "Simon Price" and foaf:workplaceHomepage of "http://www.bristol.ac.uk". Our metrics aim to capture this common-sense intuition and express any associated uncertainty in the metric value through the weighted combination of individual similarity metrics into a single overall metric.

IFPs can be accommodated and exploited where available by assigning their similarity metric (e.g. a string comparison predicate) a near maximal confidence weighting. However, depending on how the weights are combined, such a scheme could still see an IFP-based metric masked by a less reliable metric (e.g. based on *name*). For this reason, we use a "defaulting list" of procedures depending on the availability of data, both from internal and external sources, and each procedure has a confidence associated with it, which decreases for each successive procedure.

Implementation

After investigation we concluded that it was impractical to calculate the similarity metrics directly from the raw FOAF RDF triples. There are many ways to express the same thing in FOAF and some properties, such as those relating to a person's name, are not independent of each other. We decided to normalise and clean the data into our own intermediate 'internal' representation first. Besides, if this pre-processing were not done then normalisation and cleaning would have to be performed within the similarity functions themselves, considerably adding to their complexity. Given the highly structured nature of FOAF data we opted for an object-orientated representation, preferring strongly typed accessor and similarity methods for computational convenience. This intermediate representation is implemented as an SWI-Prolog module that provides a cleaned and normalised object-orientated view of the raw RDF triples. In the current implementation this is a view rather than a true pre-processing of the data and so has advantages in terms of currency but at the expense of increased computation.

Our test data is based on aggregated FOAF files from the Web collected using a purpose-written RDF crawler in SWI-Prolog. Clearly this is not representative of the fragmented partial descriptions that we are interested in; rather it is what we might expect to be the end-result of a successful smushing of partial data into more complete descriptions. Therefore, we artificially fragment the data using software, removing the IFP and other selective properties in a controlled way. This gives us precise control over the distribution and range of properties in FOAF fragments. At the same time it avoids the need to

manually label positive and negative examples of smushing.

Future Work

At the time of writing, we are continuing to develop the implementation to include a broad range of similarity metrics for comparing heterogeneous values. We plan to use the semi-artifically generated datasets as training data for machine learning algorithms in order to learn metric weights for combining into the single overall metric and to determine the threshold for smushing two descriptions.

References

[1] Dan Brickley, Libby Miller, *FOAF Vocabulary Specification*, http://xmlns.com/foaf/0.1/

[2] Eric Miller et al, *W3C Semantic Web Activity*, http://www.w3.org/2001/sw/

[3] Frank Manola, Eric Miller, eds., *Resource Description Framework (RDF)*, http://www.w3.org/RDF/

[4] Eric Miller, Jim Hendler eds., *Web Ontology Language (OWL)*, http://www.w3.org/2004/OWL/

[5] Dan Brickley, *Smushing*, http://rdfweb.org/topic/Smushing

[6] Angela Piccini, *PARIP*, http://www.bris.ac.uk/parip

[7] *TypePad*, http://www.typepad.com/